<Contributed Chair> Proteome Informatics (SGI Japan)

Contributed Chair
- Proteome Informatics (SGI Japan) -

http://www.bic.kyoto-u.ac.jp/proteome/index.html

Vis Assoc Prof
MAMITSUKA, Hiroshi
(D Sc)

Vis Instr
YAMAGUCHI, Atsuko

Scope of Research

With the advent of recently developed high-throughput experimental technologies, extremely large amounts of molecular biological data have been accumulated in these few years. With increasing the size of the data, both time- and space-efficient computational approaches have been strongly required to analyze the large-sized data sets. The primary objective of the laboratory is to establish and develop new computationally efficient methods and algorithms that allow us to better understand biologically important disciplines, such as rules, hypotheses, models and knowledge representations, from the vast amount of data of genomics and proteomics. The secondary objective is to implement the techniques developed by the laboratory as in softwares which will be used in molecular biology and related fields, such as (bio)chemistry, pharmacology and medical science. Research theme in the laboratory focuses on the issues related to proteins, with particular emphasis on protein-protein and protein-ligand interactions.

Research Activities (Year 2002)

Presentations

Dynamic Experimental Design Methodology Based on Query Learning and its Applications to Prediction of MHC Class I Binding Peptides, Abe N (IBM Watson Res. Center), Udaka K (Kyoto U), Mamitsuka H, Nakaseko Y (Kyoto U), Post-Genome Knowledge Discovery, Workshop on Protein Interactions and Clinical Data Analysis, Singapore, 28 May.

Iteratively Selecting Feature Subsets for Mining from High-Dimensional Databases, Mamitsuka H, Proceedings of the Sixth European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD02), Finland, 21 Aug.

Tree-width of chemical compounds in molecular biology

Building a method for measuring a quantitatively-significant similarity score between two arbitrary given compounds is an important issue in recent chemoinformatics and bioinformatics. If we obtain the similarity score within a relatively short time, it would be useful in a variety of both scientific and engineering applications related to chemistry and molecular biology. For tackling the issue, we deal with (the structure of) a given chemical compound as a 'molecular graph' and consider the usage of chemical and biological properties of the compound. As the first step of capturing the properties, we examined the 'tree-width', a measure indicating the complexity of a given graph, of chemical compounds particularly found in the context of molecular biology. The tree-width takes an integer in the range of *1* to *N-1* for a graph with *N* nodes and increases as with increasing the complexity of the graph. More concretely, when a graph with *N* nodes is given, the tree-width is one if it is a tree, and the tree-width is *N-1* if it is a complete graph. Figure 1 shows the schematic diagram representing the concept of the tree-width. The tree-width is a key concept in our study, because if the tree-width of a given set of graphs is bounded, it is known that for the set of graphs, a number of graph-related (NP-hard) problems, which are difficult to compute, can be solved within a relatively short (polynomial) computation time.

We obtained 9,712 chemical compounds from the LIGAND database [1] and computed the tree-width of each compound using an algorithm proposed by Matouseek and Thomas in 1991 [2]. As shown in Figure 2, in all of the 9,712 cases except only one, the tree-width of a compound is an integer in the range of one to three, and the tree-width of the one exception is four. Experimental results show that the compounds whose tree-width is three or four are limited to some particular structures. For example, most of the compounds, whose tree-width is three, are Heme, which is shown in Figure 3. Figure 4 shows the structure of the compound whose tree-width is four. From the results, we conclude that the tree-width of a chemical compound in molecular biology is all relatively small, and we will be able to develop a time-efficient graph-theoretic algorithm for dealing with the compounds by using their biological and chemical properties.

1. Goto S, Okuno Y, Hattori M, Nishioka T and Kanehisa M: LIGAND: Database of Chemical Compounds and Reactions in Biological Pathways, *Nucleic Acids Res.*, **30**, 402-404 (2002).
2. Matouseek J, Thomas R: Algorithms Finding Tree-Decompositions of Graphs, *J. Algorithms*, **12(1)**, 1-22 (1991).

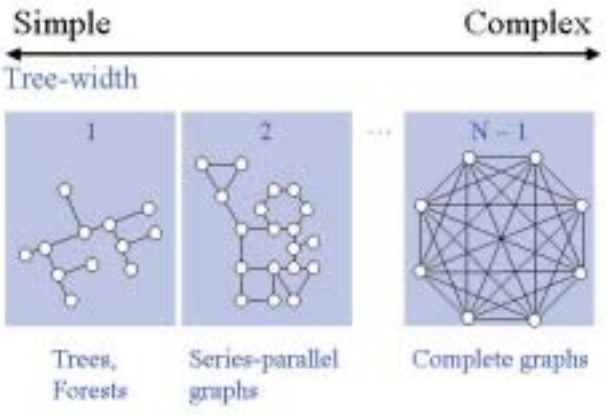

Figure 1. Tree-width: the complexity of a graph

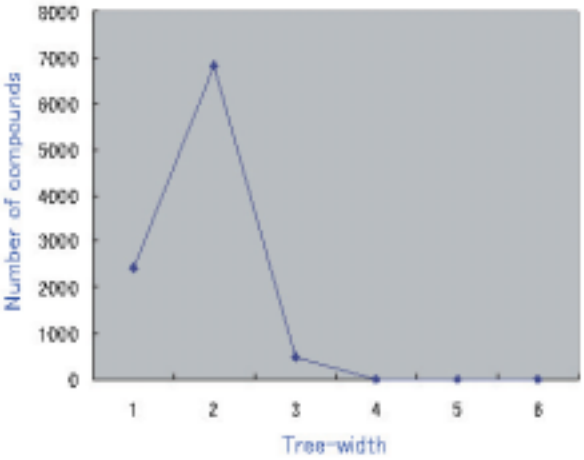

Figure 2. Distribution of tree-width of chemical compounds in LIGAND database

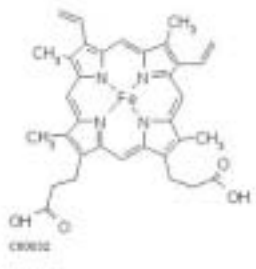

Figure 3. Heme (Tree-width: 3)

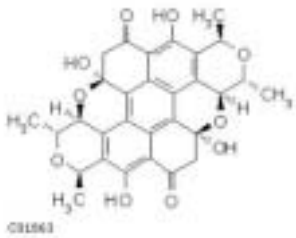

Figure 4. Xanthoaphin (Tree-width: 4)